
Proceedings Paper:
Deshmukh, N.C., Mhaske, S.S., Chandra, L.S. et al. (2026) Bias in recruitment systems utilizing large language models. In: ICAAI '25: Proceedings of the 2025 9th International Conference on Advances in Artificial Intelligence. ICAAI 2025: 2025 9th International Conference on Advances in Artificial Intelligence, 14-16 Nov 2025, Manchester, UK. Association for Computing Machinery (ACM), pp. 126-131. ISBN: 9798400721045.

https://doi.org/10.1145/3787279.3787300

Bias in Recruitment Systems Utilizing Large Language Models

Nupur Chandrashekhar Deshmukh*
Institute for Web Science and Technologies
University of Koblenz
Koblenz, Rheinland-Pfalz, Germany
nupurdeshmukh@uni-koblenz.de

Sharayu Sunil Mhaske*
Institute for Web Science and Technologies
University of Koblenz
Koblenz, Rheinland-Pfalz, Germany
sharayumhaske22@uni-koblenz.de

Lakshmi Soujanya Chandra*
Institute for Web Science and Technologies
University of Koblenz
Koblenz, Rheinland-Pfalz, Germany
soujanyachandra@uni-koblenz.de

Jayani Rachapudi*
Institute for Web Science and Technologies
University of Koblenz
Koblenz, Rheinland-Pfalz, Germany
jayanirachapudi@uni-koblenz.de

Tai Le Quy
Institute for Web Science and Technologies
University of Koblenz
Koblenz, Rheinland-Pfalz, Germany
tailequy@uni-koblenz.de

Frank Hopfgartner
Institute for Web Science and Technologies
University of Koblenz
Koblenz, Rheinland-Pfalz, Germany
University of Sheffield
Sheffield, United Kingdom
hopfgartner@uni-koblenz.de

Abstract

AI-powered recruitment systems are becoming increasingly common, offering greater efficiency in hiring processes. However, these systems may unintentionally perpetuate biases, leading to unfair hiring decisions. This study investigates various types of biases such as gender, racial, and age in recruitment systems using several large language models (LLMs) including BERT, GPT-2, and GPT-Neo. Bias in AI-driven recruitment models is assessed using the Word Embedding Association Test (WEAT) metric. The experimental results reveal varying levels of bias across LLMs and significant biases in gender and racial associations, highlighting potential disparities in AI-driven hiring recommendations. To this end, this paper underscores the need for bias mitigation strategies in AI-based recruitment tools and advocates for transparent and equitable hiring practices. The source code is publicly available at https://github.com/soujanyaachandra/Recruitment_Bias_Analysis.

CCS Concepts

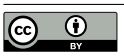
• **Computing methodologies** → **Natural language generation**.

Keywords

bias, fairness, large language models, recruitment systems

ACM Reference Format:
Nupur Chandrashekhar Deshmukh, Sharayu Sunil Mhaske, Lakshmi Soujanya Chandra, Jayani Rachapudi, Tai Le Quy, and Frank Hopfgartner. 2025. Bias in Recruitment Systems Utilizing Large Language Models. In *2025 9th International Conference on Advances in Artificial Intelligence (ICAAI 2025), November 14–16, 2025, Manchester, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3787279.3787300

*These authors contributed equally.



ICAAI 2025, Manchester, United Kingdom

ACM ISBN 979-8-4007-2104-5/25/11
https://doi.org/10.1145/3787279.3787300

1 Introduction

Artificial intelligence (AI) is becoming increasingly prevalent in recruitment processes, offering efficiency and scalability in hiring decisions. However, the use of AI-driven hiring algorithms raises serious ethical concerns, particularly regarding bias and equity [2, 11]. The integration of AI in hiring processes has shown promise in reducing human error, processing large volumes of applications quickly, and potentially mitigating unconscious biases. However, this potential is threatened by the reality of biased AI systems, which can be derived from historical data used to train AI models, perpetuating social stereotypes and leading to unfair hiring decisions [9].

The implications of biased AI recruitment systems are profound. Qualified candidates from underrepresented groups can be systematically overlooked, perpetuating a lack of diversity within organizations and hindering overall innovation [8]. These biases can reinforce existing inequalities in the job market and potentially violate anti-discrimination laws [8]. The increasing reliance on AI in recruitment demands immediate attention, as unchecked biases can exacerbate existing societal inequalities and undermine the principles of fair employment [2, 8].

Understanding and mitigating these biases is crucial to ensuring fair hiring practices in the age of AI. Recent studies have shown that AI systems can exhibit gender and racial biases, leading to discriminatory outcomes in hiring processes [8]. For example, research has shown that several AI-powered resume screening tools may favor candidates with names typically associated with certain ethnic groups, potentially disadvantaging equally qualified applicants from other backgrounds [2].

To address the aforementioned challenges, this paper will explore several key aspects of bias in AI-driven recruitment models. We will examine the extent of bias in models using LLMs such as BERT, GPT-2, and GPT-Neo, which have shown promising results in natural

language processing tasks, but may also inherit biases present in their training data. To quantify bias in LLMs-based systems, we use the Word Embedding Association Test (WEAT) measure.

The rest of our paper is structured as follows: Section 2 overviews the related work. The methodology is presented in Section 3. Section 4 presents the details of our experiments on various LLMs. Finally, the conclusion and outlook are summarized in Section 5.

2 Related work

AI is increasingly applied in HR for tasks such as attrition prediction, chatbot-based services, and resume screening with many tools such as Google Hire, HireVue, etc. [6]. LLMs have been used in recruitment systems to analyze candidates' resumes. Vijayalakshmi et al. [13] proposed a methodology that leverages ChatGPT-3.5 Turbo, sentence transformers, and a vector database to efficiently analyze and compare resumes with job descriptions, performing skill extraction, matching, and scoring. Yerkebulan et al. [14] utilized the BERT-based model to automate the recruitment process through resume analysis. Gan et al. [3] developed an LLMs-based agent framework for resume screening, aimed at enhancing efficiency and time management in recruitment processes.

Njoto et al. [7] conducted a simulation of a real-world recruitment setting to examine how algorithms may introduce human bias (gender bias) into hiring decisions. Hofeditz et al. [4] examined the impact of AI-generated candidate recommendations on human hiring decisions and explored the moderating role of Explainable AI. By integrating Gibson's direct perception theory and Gregory's indirect perception theory, Soleimani et al. [9] bridged psychological, information systems, and human resource management (HRM) perspectives to advance the understanding of decision-making biases in AI. Recently, with the boom of LLMs, GPT has been leveraged [5] to generate high-quality, diverse synthetic data for retraining AI systems aimed at identifying and mitigating gender bias in the recruitment process.

3 Methodology

3.1 Dataset

We use the **Recruitment dataset**[1] which consists of 10,000 job applicants, each represented by 9 features that encapsulate demographic information such as age, race, gender, job-related details, and textual descriptions, including resumes and job descriptions. These features enable for an in-depth analysis of recruitment bias using LLMs. The dataset contains an outcome variable "Best Match" which indicates whether the candidate's resume aligns with the job description (0 = "No Match", 1 = "Match"). The age distribution ranged from 25 to 55 years (mean = 40, standard deviation = 8.95). The dataset maintains a balanced gender distribution: 5,059 males (50.6 %) and 4,941 females (49.4%). The racial composition is evenly distributed, with 33.6 % Mongoloid/Asian, 33.2% White/Caucasian, and 33.2 % Negroid/Black applicants. The dataset covers 51 unique job roles, with notable concentrations in the technical, managerial, and healthcare professions.

We perform the preprocessing tasks on the dataset. All text was converted to lowercase and non-alphabetic characters, and special symbols were removed using regular expressions. Then, the text was tokenized and lemmatized, the stop words (English) were removed using NLTK's toolkit[2]. In addition, the pre-processed resumes and job descriptions were combined into a single list for the subsequent feature extraction phase. In the next step, TF-IDF was initially used to extract a list of the most popular keywords from the preprocessed text. These keywords were further refined using cosine similarity with predefined STEM and Non-STEM reference word lists. In particular, STEM reference includes "algorithm", "neural", "processor", "dataset", "AI", "ML", "deep", "network", "programming", "software", "hardware", "databases", "cybersecurity", "engineering", "science", "physics", "mathematics", "robotics", "biotechnology", "statistics", "machine learning", "data science", "cloud computing", "computer vision", "natural language processing", "big data", "automation", "genomics", "bioinformatics", "quantum computing", "electrical engineering". Non-STEM reference consists of "management", "leadership", "marketing", "sales", "communication", "teamwork", "customer", "business", "strategy", "collaboration", "writing", "counseling", "creativity", "organization", "planning", "support", "human resources", "social work", "public relations", "event planning", "journalism", "advertising", "psychology", "education", "training", "hospitality", "tourism", "real estate". TF-IDF was applied to the combined text data to identify the most important words. A stop list was used to filter out common words that do not provide a meaningful classification. Furthermore, we apply the frequency encoding on "Job Applicant Name", "Ethnicity", and "Job roles" to convert these categorical attributes. In particular, "Ethnicity" and "Job Roles" are frequency encoded using the observed frequency of that category in the data, allowing the model to identify trends correlated with ethnicity and/or job roles without implying ranking or ordering of the categories. Furthermore, we apply RobustScaler to numerical features, since it is more robust in cases with extreme values. Because skewed data can affect model performance [10], we calculate the skewness for each numerical feature. There are two skewed features that require transformation: "Job Applicant Name" (skewness: 0.1426) and "Job Roles" (skewness: -0.1150). We employ the Power Transformer with the Yeo-Johnson transformation. The updated skewness are 0.0360 and 0.0420 for those features, respectively.

3.2 Word Embedding Association Test (WEAT)

Word Embedding Association Test (WEAT) [1] is a common method used to measure bias in word embeddings. The WEAT metric works by computing the cosine similarity between different sets of words. The key idea is to compare the similarity of target words with attribute words. This is done by implementing the WEAT effect size. Let the target sets (X, Y) be the concepts you are testing for association and the attribute sets (A, B) be the groups you are comparing. We denote $s(w, A, B)$ as the association score of a word w with the attribute sets A and B.

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(w, a) - \frac{1}{|B|} \sum_{b \in B} \cos(w, b) \qquad (1)$$

[1] https://www.kaggle.com/datasets/surendra365/recruitement-dataset

[2] https://www.nltk.org/

Then the WEAT effect size d is computed as follows:

$$d = \frac{\text{mean}_{x \in X}\, s(x, A, B) - \text{mean}_{y \in Y}\, s(y, A, B)}{\text{std dev}_{w \in X \cup Y}\, s(w, A, B)} \quad (2)$$

The effect size d is standardized, so it's not influenced by the scale of the data [12]. A positive value of d indicates that X is more strongly associated with A, and a negative value of d indicates a stronger association with B.

4 Evaluation

4.1 Experimental Setups

We extract the two target words, STEM and Non-STEM, from the dataset because they better capture the distinction between the unique job roles. The implementation involved multiple steps: data preprocessing, keyword extraction, word embedding extraction, and bias quantification. The same process is applied across multiple LLMs. For each LLM, the corresponding tokenizer and model are loaded using the Hugging Face library. In particular, BERT[3], GPT-2[4], and GPT-Neo[5] are used for embedding generation. Keywords were assigned to the category STEM or Non-STEM with the highest average cosine similarity, above a similarity threshold (0.3). The similarity threshold of 0.3 was chosen for better word filtering and to avoid noise. Each word is then assigned to STEM or Non-STEM based on the higher similarity score, forming two final lists, i.e., STEM keywords and Non-STEM keywords. Since the embeddings are model-specific, the STEM and Non-STEM keyword lists will vary from model to model. These sets were preserved for bias detection using WEAT.

4.2 Experimental Results

4.2.1 Correlation Analysis. The correlation heatmap (Figure 1) reveals notable patterns that suggest potential biases in the job selection process. There is a clear correlation between "gender" and the "Best Match" class label. This pattern indicates a tendency to favor male candidates in the selection process, warranting further investigation. Additionally, a strong correlation between "Job Applicant Name" and "Race" suggests the possibility of racial bias, potentially arising from name-based assumptions influencing hiring decisions. In contrast, "Age" and "Ethnicity" exhibit weak correlations with the "Best Match" outcome, implying that they may not play a significant role in selection. However, the presence of more subtle or indirect biases cannot be ruled out.

4.2.2 Biases Analysis. The WEAT (Eq. 2) is used to measure different types of biases-gender, racial, and age-within the contexts of STEM and Non-STEM fields.

Gender bias. Figure 2 illustrates the gender bias detected using the WEAT effect sizes across three language models: BERT, GPT-2, and GPT-Neo. A negative WEAT effect size for "STEM vs. Male" suggests that STEM fields are less associated with males, while a positive effect size for "STEM vs. Female" indicates that STEM fields are more associated with females. Interestingly, all three LLMs exhibit this reversed bias, contradicting the common societal

Figure 1: Correlation heatmap

stereotype that STEM is more male-associated. In particular, the BERT model demonstrates the strongest gender bias, associating STEM more with females and Non-STEM with males. A similar trend is observed in the results of the GPT-2 and GPT-Neo, but GPT-Neo model has the weakest bias among the three models. For Non-STEM fields, the pattern is reversed: Non-STEM fields are less associated with males and more associated with females. These findings confirm that while biases exist, they do not align with traditional societal stereotypes, which warrants further analysis of how these models internalize and reflect gender associations.

First name bias (Race associations). Figure 3 visualizes the first name bias detected across the LLMs. We examine how strongly different LLMs associate "White", "Black", and "Asian" names with the general concept of a first name. In detail, the BERT model exhibits a strong positive association between "White" names and general first names, indicating a bias where "White" names are perceived as more prototypical first names compared to "Black" or "Asian" names. The GPT-2 and GPT-Neo models display a negative association between "White" names and general names, suggesting a bias where "Black" and "Asian" names are more strongly associated with the concept of a first name. Across all LLMs, the bias between "Black" and "Asian" names is relatively low, meaning these two groups are treated more similarly compared to their association with "White" names. These findings suggest that language models encode varying biases in name associations, which may have significant implications for applications involving name-based predictions and identity recognition.

Racial bias (STEM/Non-STEM associations). The BERT model shows a strong positive association between "White" names and STEM, reinforcing the stereotype that "White" individuals are more closely linked to STEM careers compared to "Black" or "Asian" individuals (Figure 4). In contrast, the GPT-2 and GPT-Neo models have a negative association between "White" names and STEM, instead showing a bias where "Black" and "Asian" names are more strongly

[3]https://huggingface.co/google-bert/bert-base-uncased
[4]https://huggingface.co/openai-community/gpt2-medium
[5]https://huggingface.co/EleutherAI/gpt-neo-125m

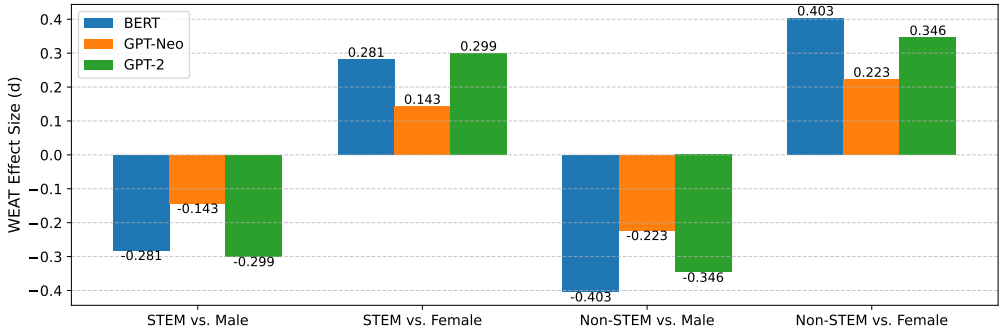


Figure 2: Comparing gender bias across different LLMs

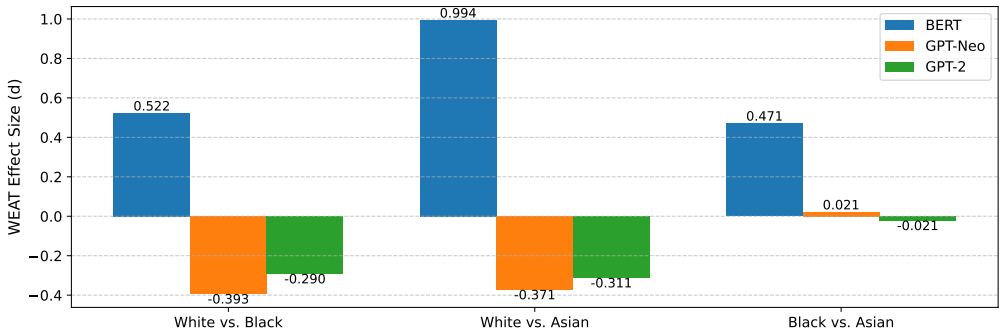


Figure 3: Comparing first name bias (race associations) across different LLMs

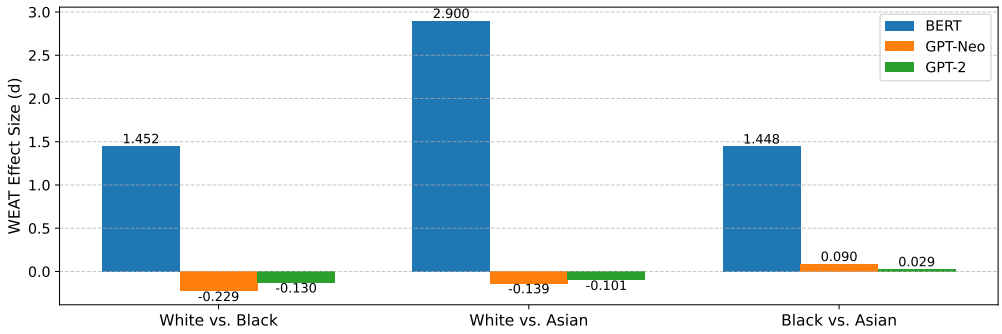


Figure 4: Comparing racial bias (STEM/Non-STEM associations) across different LLMs

associated with STEM fields—though the effect sizes are smaller than BERT's "White-STEM" bias. In all LLMs, the bias between "Black" and "Asian" names in the STEM associations is relatively low, indicating that these groups are treated more similarly compared to their association with "White" names.

"Best Match" bias (STEM/Non-STEM associations). Figure 5 presents the "best match" bias analysis, examining how strongly LLMs associate "White","Black", and "Asian" names with being the "best match" for STEM fields compared to Non-STEM fields. In particular, the BERT model shows a strong positive bias, associating "White" individuals as the "best match" for STEM careers, reinforcing a pro-White STEM preference. However, the GPT-2 and GPT-Neo models exhibit a negative bias, instead suggesting that "Black" and "Asian" individuals are more strongly linked as the "best match" for STEM—though with smaller effect sizes than BERT's "White-STEM" association. The bias strength varies significantly between models, with BERT encoding the strongest racial bias, while GPT-2 and GPT-Neo show weaker but inverse tendencies.

Age bias (STEM/Non-STEM associations). Age bias analysis in STEM/Non-STEM associations reveals minimal systematic bias across all models, as shown in Figure 6. The observed effects are substantially weaker than gender and racial biases and the patterns remain consistent across BERT, GPT-2, and GPT-Neo architectures.

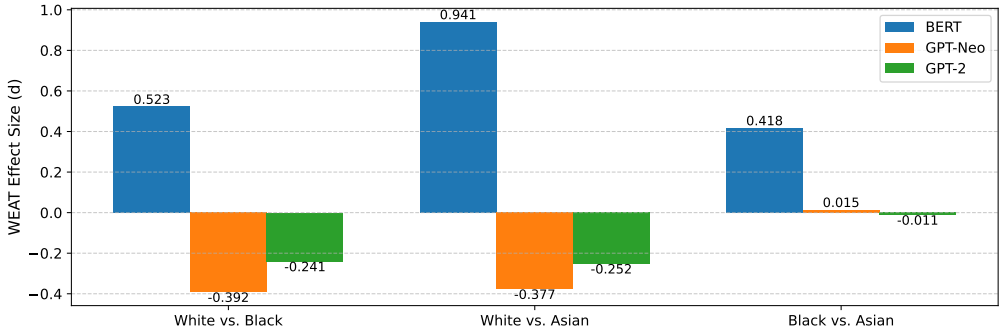


Figure 5: Comparing "Best match" bias (STEM/Non-STEM associations) across different LLMs

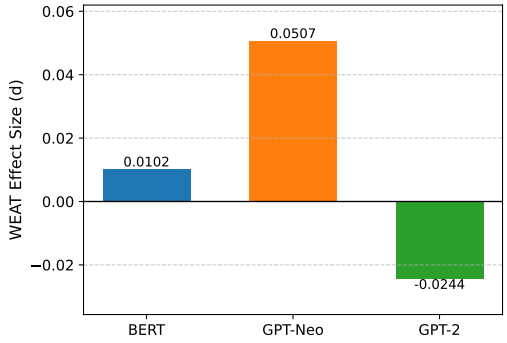


Figure 6: Comparing age bias (STEM/Non-STEM associations) across different LLMs

In general, BERT exhibits the most substantial biases across most categories, particularly in racial biases related to name associations and STEM fields, displaying a prominent pro-White tendency. This is likely due to BERT's training in a vast corpus of historical data, which inherently reflects and amplifies the existing societal biases prevalent during those periods. In contrast, GPT-Neo and GPT-2 generally show weaker biases and, in some cases, indicating a pro-Black/Asian association. This reversal could stem from several factors: potentially, these autoregressive models might have been trained on more recent, diverse datasets that reflect evolving societal norms, or they might inherently process and contextualize information differently, leading to varied bias representations. All models show reverse gender bias, associating STEM with female names, which is a surprising result that warrants further investigation. Moreover, all models demonstrate minimal age bias. These variations highlight the differences in how these models learn and represent societal biases from their training data, emphasizing the necessity for careful examination and mitigation of biases in LLMs.

5 Conclusions and Outlook

In this paper, we have investigated bias in recruitment systems utilizing LLMs, focusing on BERT, GPT-2, and GPT-Neo. Our findings reveal that all LLMs exhibit some forms of bias related to gender and race, although the direction and magnitude vary across models. This confirms that bias remains an intrinsic challenge in AI-driven recruitment systems. These results underscore the need for comprehensive bias detection and mitigation strategies in AI-powered recruitment tools. Although LLMs offer powerful capabilities for processing and analyzing job applications, their deployment must be accompanied by rigorous fairness assessments and interventions to ensure equitable outcomes for all demographic groups. In future work, we plan to investigate how biases interact across multiple protected attributes (e.g., race and gender) to produce compounded effects for certain groups. In addition, we aim to explore bias measures including outcome-based measures and causal frameworks, and to develop tools that make biases transparent and interpretable to non-technical stakeholders.

References

[1] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. doi:10.1126/science.aal4230

[2] Zhisheng Chen. 2023. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and social sciences communications* 10, 1 (2023), 1–12. doi:10.1057/s41599-023-02079-x

[3] Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2024. Application of LLM agents in recruitment: a novel framework for automated resume screening. *Journal of Information Processing* 32 (2024), 881–893. doi:10.2197/ipsjjip.32.881

[4] Lennart Hofeditz, Sünje Clausen, Alexander Rieß, Milad Mirbabaie, and Stefan Stieglitz. 2022. Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring. *Electronic Markets* 32, 4 (2022), 2207–2233. doi:10.1007/s12525-022-00600-9

[5] Donghyeok Lee, Rajesh R Jaiswal, and Adrian Byrne. 2024. Utilising Synthetic Data from LLM for Gender Bias Detection and Mitigation in Recruitment Systems. In *HCAIep '24*. 57–57. doi:10.1145/3701268.3701285

[6] Dena F Mujtaba and Nihar R Mahapatra. 2019. Ethical considerations in AI-based recruitment. In *IEEE ISTAS 2019*. IEEE, 1–7. doi:10.1109/ISTAS48451.2019.8937920

[7] Sheilla Njoto, Marc Cheong, Reeva Lederman, Aidan McLoughney, Leah Ruppanner, and Anthony Wirth. 2022. Gender bias in AI recruitment systems: A sociological-and data science-based case study. In *IEEE ISTAS 2022*, Vol. 1. IEEE, 1–7. doi:10.1109/ISTAS55053.2022.10227106

[8] Päivi Seppälä and Magdalena Małecka. 2024. AI and discriminative decisions in recruitment: Challenging the core assumptions. *Big Data & Society* 11, 1 (2024), 20539517241235872. doi:10.1177/20539517241235872

[9] Melika Soleimani, Ali Intezari, James Arrowsmith, David J Pauleen, and Nazim Taskin. 2025. Reducing AI bias in recruitment and selection: an integrative grounded approach. *The International Journal of Human Resource Management* (2025), 1–36. doi:10.1080/09585192.2025.2480617

[10] PN SubbaNarasimha, B Arinze, and M Anandarajan. 2000. The predictive accuracy of artificial neural networks and multiple regression in the case of skewed data: Exploration of some issues. *Expert systems with Applications* 19, 2 (2000), 117–123. doi:10.1016/S0957-4174(00)00026-9

[11] Nicholas Tilmes. 2022. Disability, fairness, and algorithmic bias in AI recruitment. *Ethics and Information Technology* 24, 2 (2022), 21. doi:10.1007/s10676-022-09633-2

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. *Advances in Neural Information Processing Systems* 30 (2017).

[13] V Vijayalakshmi, A Ananya, and MU Sharanya Avadhani. 2024. Optimization of HR Recruitment Process using Large Language Model (LLM). In *ICICEC 2024*. IEEE, 1–5. doi:10.1109/ICICEC62498.2024.10808420

[14] Alimzhan Yerkebulan, Madina Mansurova, Assel Abdildayeva, and Olzhas Sharip. 2025. Research on the Application of Large Language Models (LLM) for Improving Recruitment Efficiency and Accuracy. In *ACIIDS 2025*. Springer, 331–344. doi:10.1007/978-981-96-6008-7_24